Volume 3, Number 3, 2026
Pages 1569-1594
Bioscience Insights
ISSN: 3080-3489 (Print)
ISSN: 3080-3497 (Online)
BioSciInsights.com

AI-DRIVEN PRECISION MENTAL HEALTH FRAMEWORK USING MULTIMODAL BEHAVIORAL ASSESSMENT, EXPLAINABLE MACHINE LEARNING, AND PERSONALIZED PSYCHOLOGICAL INTERVENTION PREDICTION FOR EARLY DETECTION AND PERSONALIZED TREATMENT OF MENTAL DISORDERS

Taimour Mushtaq[*1], Muhammad Hammad Gohar[2], Irfa Masood[3], Esha Yasin[4], Jasika Patras[5], Alisha Rashid[6]

*[*1]Department of Computer Science, Federal Urdu University of Arts, Science and Technology, Islamabad, Pakistan.*
[2]Department of Sciences, Broomfield Hall School, Peshawar, Pakistan.
[3] Independent Researcher, Islamabad, Pakistan.
[4]Department of Public Health, Babeş-Bolyai University, Cluj-Napoca, Romania.
[5]Department of Nursing, Iqra University, North Campus, Karachi, Pakistan.
[6]Independent Researcher, Rahim Yar Khan, Pakistan.

[*1]taimourmushtaq.cs@fuuast.edu.pk, [2]hammadgohar103@gmail.com, [3]irfamasood962@gmail.com, [4]ishayasin44@gmail.com, [5]jasika.patras@iqra.edu.pk, [6]alisharashid2020@gmail.com

https://doi.org/10.5281/zenodo.24105959

Article Information

Article History

Received:	June 19, 2026
Revised:	August 19, 2026
Accepted:	September 07, 2026
Available Online:	November 10, 2026

Keywords:
Precision mental health; digital phenotyping; multimodal behavioral assessment; explainable machine learning; early mental disorder detection; personalized psychological intervention; treatment response prediction

Copyright @Author
Corresponding Author: *
Taimour Mushtaq

Abstract

Mental disorders frequently remain undetected until symptoms become clinically significant, while conventional diagnostic approaches rely on intermittent assessments, subjective reporting, and trial-and-error treatment selection. This study proposes an AI-driven precision mental health framework integrating multimodal behavioral assessment, explainable machine learning, and personalized psychological intervention prediction to improve early identification, diagnostic differentiation, and individualized treatment matching. The framework combines four complementary data modalities: smartphone and wearable sensing, speech acoustics, linguistic behavior, and standardized self-report assessments. A reproducible simulated cohort comprising 2,400 participants from three recruitment sites was developed, including 952 healthy individuals, 645 participants with depression, 494 with anxiety, and 309 with bipolar disorder. Twenty-one behavioral and psychological features were integrated using a gradient-boosted learning architecture, while permutation-based feature attribution enhanced interpretability. Model performance was evaluated through stratified testing, five-fold cross-validation, cross-site generalization, longitudinal monitoring, subgroup analysis, and simulated randomized treatment allocation.

Experimental findings demonstrated strong discrimination across psychiatric categories, with the multimodal model achieving a macro-averaged area under the curve (AUC) of 0.960 (95% CI: 0.949–0.970), macro F1-score of 0.832, and classification accuracy of 85.0% on 600 held-out participants. Individual behavioral modalities achieved AUC values of 0.762–0.777, compared with 0.899 for self-report assessments. Cross-validation produced an AUC of 0.962 ± 0.003, while cross-site performance ranged from 0.954 to 0.973. Explainability analysis identified GAD-7 scores and mood variability as influential predictors of anxiety and bipolar disorder, respectively. Longitudinal monitoring detected 99.7% of symptomatic participants at 95% specificity, providing an average warning lead of 2.64 weeks under simulation-defined onset assumptions. Eight-week multimodal monitoring increased sensitivity to 88.1%. Personalized treatment matching improved simulated response rates from 39.1% to 53.2%, representing a 14.1-percentage-point improvement. These simulation-based findings demonstrate methodological feasibility rather than established clinical effectiveness and support prospective validation using real-world psychiatric cohorts.

INTRODUCTION

Depression, anxiety and bipolar disorder impose a burden that is both large and unevenly recognized. Global estimates of the prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 have been produced that account for the effects of the COVID-19 pandemic [1], and the World Health Organization has called for a transformation of mental health care [2]. Delay is one of the clearest symptoms of the problem. A systematic scoping review of bipolar disorder in adolescents and young adults examines the factors that influence delays in diagnosis and treatment [3]. A person who is already symptomatic may therefore wait a long time before an accurate label, and only then begin the slow process of finding a treatment that works.

Precision psychiatry has been proposed as a way to shorten that process by replacing population averages with individualized prediction. Systematic reviews of individualized prediction models examine what is needed to implement them in clinical practice [4], and psychotic disorders have been proposed as a framework for precision psychiatry more broadly [5]. The promise is attractive, but it depends on measurement. Clinical interviews and questionnaires capture a participant at a few moments, whereas the conditions of interest fluctuate over weeks. A framework that is serious about early detection needs a data stream that is dense in time and sensitive to behavior change before the person is able to report it.

Digital phenotyping offers such a stream. Systematic reviews of the digital footprint of depression [6] and of digital phenotyping for stress, anxiety and mild depression [7] have collected empirical evidence on the behavioral traces that digital devices capture in relation to depression, stress and anxiety. Yet the same literature exposes three weaknesses that motivate the present work. First, studies tend to analyze one data stream at a time, so the complementary value of sensors, speech, text and self-report is rarely measured under a common protocol. Second, the models are often opaque, even though clinicians are unlikely to act on risk scores they cannot interrogate, a concern that work on explainable artificial intelligence for mental health addresses through transparency and interpretability [9]. Third, continuous monitoring raises ethical questions about consent, data ownership and unintended inference that developers have only recently begun to formalize through structured expert consensus [8].

A fourth weakness concerns the gap between prediction and action. Detecting a disorder is not the same as choosing what to do about it. A systematic review and meta-analysis has examined machine learning prediction of depression treatment outcomes [11], and recent work has predicted individual patient outcomes under psychotherapy versus medication for major depression [12]. Few frameworks link these two ends, so that the same behavioral representation used for detection also informs the treatment choice.

Meanwhile, whether any such link generalizes across populations is uncertain. Multi-year datasets were built so that the generalization of longitudinal behavior models across cohorts could be measured [13], fairness and bias correction in depression prediction have been examined across four study populations [14], and gender fairness in mental health prediction has been addressed specifically [10].

Taken together, these gaps explain why a framework, rather than a single model, is the appropriate unit of contribution. A detector that is accurate but unexplained will not be trusted, an explained model that is evaluated on one site will not be believed to generalize, and a prediction that is never linked to a decision leaves the clinician with the original problem of choosing what to do. Each component of the framework is therefore tied to a specific weakness: the multimodal learner addresses the single-stream limitation, the explanation layer addresses opacity, the subgroup and cross-site audit addresses fairness and generalization, and the treatment head addresses the gap between prediction and action. The ethical concerns raised by continuous monitoring [8] are treated as a design constraint, which is why the framework works from summary features rather than raw audio or text and why Section 6 returns to consent and data minimization.

This paper responds to these gaps with a single integrated framework, summarized in Fig. 1. The framework takes four behavioral streams (smartphone and wearable sensing, speech acoustics, language use, and self-report with ecological momentary assessment), harmonizes them with missing-modality handling in a unified feature table, learns a calibrated multimodal risk model, attributes that risk to individual features, and then reuses the same representation for three clinical decisions: weekly early-detection alerts, differential stratification across depression, anxiety and bipolar disorder, and matching of patients to cognitive behavioral therapy or medication.

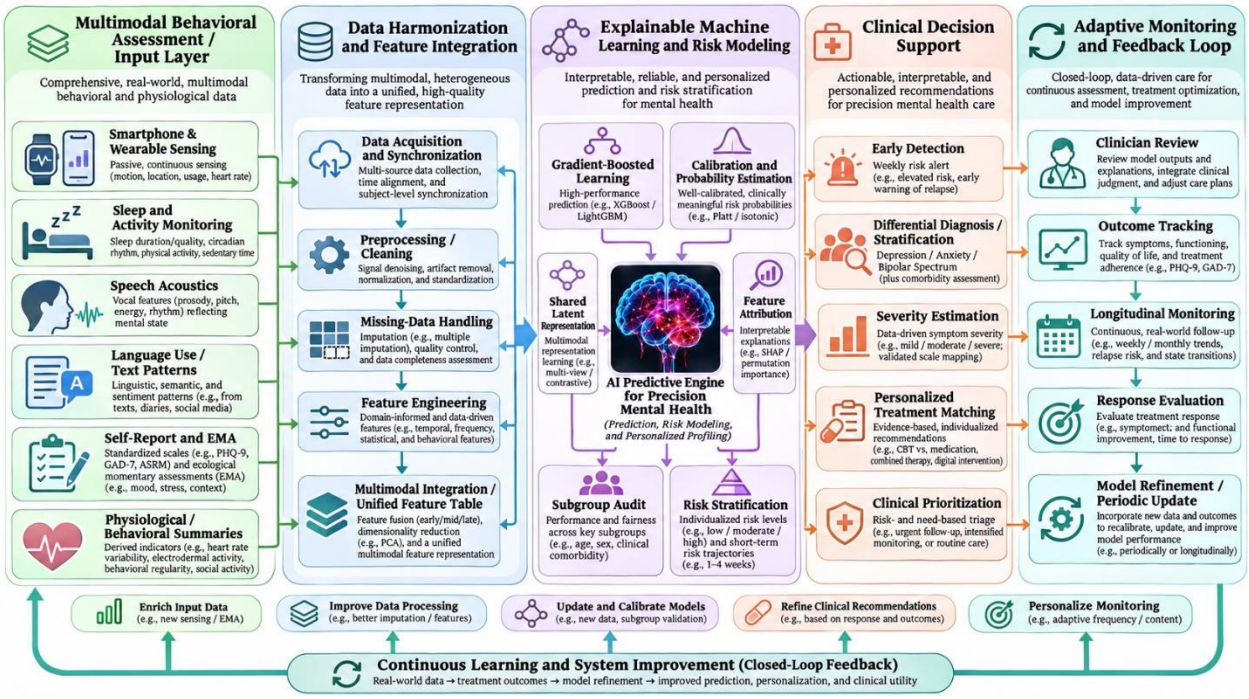


Fig. 1. Advanced AI-driven precision mental health architecture integrating multimodal behavioral data fusion, explainable predictive modelling, personalized clinical decision support, and continuous feedback-based monitoring.

Figure 1 presents the comprehensive architecture of the proposed AI-driven precision mental health framework, illustrating the interconnected processes of multimodal behavioral assessment, intelligent data harmonization, explainable machine learning, personalized clinical decision-making, and adaptive longitudinal monitoring. The framework begins with the acquisition of complementary behavioral and psychological information from smartphone and wearable sensing, sleep and activity monitoring, speech acoustics, language patterns, self-report questionnaires, and physiological summaries. These data streams

capture variations in daily activity, mobility, sleep regularity, vocal characteristics, emotional expression, mood fluctuations, and psychological symptoms. The acquired information is subsequently processed through data synchronization, noise reduction, missing-value management, feature engineering, and multimodal integration to construct a unified representation of individual behavioral and psychological characteristics.

The central analytical layer employs a calibrated gradient-boosted machine learning model to identify patterns associated with depression, anxiety, and bipolar disorder while generating individualized risk estimates and predicted treatment responses. Explainability mechanisms, including permutation-based feature attribution, support the identification of influential predictors and assessment of their contributions to disorder-specific predictions. Additional subgroup evaluations examine model consistency across demographic categories and recruitment sites, while calibration analysis assesses the reliability of predicted probabilities. The resulting predictive outputs support five interconnected clinical decision functions: early detection of emerging psychological symptoms, differential diagnostic stratification, symptom-severity estimation, personalized treatment matching, and risk-based clinical prioritization. These functions are intended to assist clinician-led assessment rather than replace professional diagnosis or treatment decisions.

The final component introduces an adaptive monitoring and feedback mechanism incorporating clinician review, treatment outcome tracking, longitudinal symptom assessment, response evaluation, and periodic model refinement. Feedback pathways connect observed outcomes with earlier data-processing and predictive-modelling stages, creating a conceptual closed-loop system for improving feature representations, reassessing risk predictions, and refining intervention recommendations. The architecture also highlights the importance of continuous performance evaluation, subgroup fairness, and clinically interpretable predictions for developing responsible precision mental health systems. Overall, the proposed framework establishes an integrated methodological foundation for combining heterogeneous behavioral evidence with explainable artificial intelligence to investigate earlier mental disorder identification and more individualized psychological interventions.

The contribution is deliberately scoped. Because access to a single real cohort that contains all four modalities, longitudinal onset information and randomized treatment assignment is rare, the framework is demonstrated on a transparent simulated cohort of 2,400 participants whose generating assumptions are stated in full. The simulation allows every claim to be reproduced and every limitation to be located, but it also means that none of the figures reported below should be read as clinical performance. The remainder of the paper is organized as follows. Section 2 reviews the evidence on each behavioral stream and on treatment prediction and locates the gaps that the framework targets. Section 3 describes the cohort, features, models and evaluation design. Section 4 reports the results and discusses them. Section 5 concludes, Section 6 sets out limitations and future work, and Section 7 lists the references.

2. Literature Review

2.1 Smartphone and wearable sensing

Smartphone sensing is the most developed strand of digital phenotyping. A PRISMA review of the empirical literature on the digital footprint of depression synthesizes how behavioral traces from phones relate to depressive symptoms [6], and a later systematic review extends the scope to stress, anxiety and mild depression [7]. At the level of individual prediction, personalized mood prediction from patterns of behavior collected with smartphones has been developed to move beyond one-size-fits-all models [15], and observational work on smartphone-based sensing features has examined how well they predict depression severity [16]. Wearables add physiological context to behavior. A systematic review and meta-analysis of wearable artificial intelligence for detecting anxiety consolidates evidence for that disorder [18], digital biomarkers have been used to build personalized relapse predictors for major depressive disorder [17], and interpretable frameworks have been developed that classify depression and its severity from activity data [19]. Taken together, this body of work establishes that passive streams carry usable signal, but it also leaves the question of how sensing combines with speech, language and self-report largely open.

Two methodological themes run through this strand. The first is the contrast between population-level and person-specific models. Personalized mood prediction from smartphone behavior [15] and

personalized relapse prediction from digital biomarkers [17] both reflect the view that the same behavioral value can mean different things for different people, which makes within-person baselines attractive and also makes models harder to transfer between individuals. The second theme is generalization. Multi-year datasets were assembled so that the transfer of longitudinal behavior models between cohorts could be measured [13], which suggests that single-cohort results cannot be assumed to carry over. For a framework intended for clinical use, both themes imply that performance should be reported under cross-site testing and that the limits of population-level features should be stated plainly.

2.2 Speech and acoustic analysis

Speech carries information about psychomotor state, affect and cognition that other streams cannot reach. Smartphone-based, text-dependent speech signals have been used with deep convolutional networks to detect depression automatically [20], and automated speech analysis has been validated for risk detection across depression, anxiety, insomnia and fatigue within one algorithm development study [21]. The practical attraction is that a short voice sample can be collected on the same device that supplies passive data. The limitation is that speech studies typically report detection of a single condition and rarely include a comparison with the other behavioral streams, so the incremental value of voice over sensing and self-report remains difficult to judge.

The speech literature also illustrates a trade-off between rich and practical measurement. Deep models applied to smartphone speech signals [20] can learn directly from the recording, whereas analyses designed to cover several conditions at once, such as depression, anxiety, insomnia and fatigue [21], must separate signals that overlap across conditions. For a multi-class monitoring framework the second setting is the more relevant, because the clinical question is rarely whether a person is unwell in general but which condition best explains the pattern. This is the reason the present framework treats speech as a source of compact acoustic features, such as pitch variability, speech rate, pause ratio and jitter, whose contribution can be separated from that of the other streams.

2.3 Multimodal fusion

Multimodal systems combine streams in an attempt to recover what each misses alone. A macro-micro personalization framework for multimodal-multitask learning has been proposed for mental health monitoring [22], depression diagnosis from patient interviews has been framed as a multimodal machine learning problem [23], and facial multimodal data have been used for diagnosis [24]. Deep models that jointly process video, audio and text have also been developed to predict depression [25], and behavioral and voice data have been combined for early detection of mental health disorders with conventional machine learning [26]. These studies make a persuasive case for fusion, but most use interview or video modalities that are collected in a clinical or laboratory setting, which limits their usefulness for continuous monitoring in daily life. They also tend to optimize detection in isolation, without carrying the representation forward to explanation or treatment choice.

Two broad strategies are available for combining streams. Early fusion concatenates features from all streams into a single input, which allows interactions between streams to be learned but requires a common handling of missing data. Late fusion trains a model per stream and combines their outputs, which tolerates missing streams but cannot exploit interactions. The framework proposed here adopts early fusion with native handling of missing values, because it keeps one representation for every downstream task, and it compares that choice against single-modality and partial-fusion models in an explicit ablation. The reviewed fusion studies [22]–[26] differ in architecture, but they share the property that the benefit of each added stream is rarely isolated, which is the measurement that Fig. 5 and Fig. 9 provide.

2.4 Language models and text

Language is the newest modality. Large language models have been used to detect depression from user-generated diary text as a screening approach [27], and a systematic review of large language models for mental health applications has surveyed how they are being used [28]. For a monitoring framework, text offers a low-burden stream, because diary or message content can be summarized into compact features

such as affect words, first-person pronouns and sentiment variability. The same properties raise privacy concerns that are discussed in Section 2.7, and the evidence base is still thin on how language features behave when combined with sensing and self-report under a shared evaluation.

2.5 Bipolar spectrum and differential diagnosis

Most digital phenotyping work concentrates on depression, but mood disorders are differential problems. A systematic review and meta-analysis of machine learning for the diagnostic accuracy of bipolar disorder reviews the diagnostic accuracy of existing classifiers [31], while a cohort protocol for relapse prediction in bipolar disorder proposes passive sensing and deep anomaly detection to learn patient-specific behavior [30]. The clinical stakes of getting this differential right are underlined by the evidence on diagnostic delay in young people with bipolar disorder [3]. A monitoring framework that only detects depression risks mislabeling the cases for which an accurate differential matters most.

2.6 Treatment response prediction

Prediction of treatment outcome is the step that converts a diagnosis into a plan. A systematic review and meta-analysis of machine learning in the prediction of depression treatment outcomes [11] and a more recent systematic review of machine learning for therapeutic outcome prediction in major depressive disorder [29] describe a maturing field. Newer work asks the comparative question directly by predicting individual patient outcomes under psychotherapy versus medication for major depression [12]. What this literature rarely does is to show that the representation used for detection can also guide treatment selection, which is the integration that the present framework attempts to test in a controlled setting.

2.7 Explainability, fairness, generalization and ethics

Transparency and interpretability are repeatedly named as preconditions for clinical adoption of artificial intelligence in mental health [9]. Fairness research shows that a model with good average performance can still fail for specific groups, whether the grouping is gender [10] or the study population in which depression prediction is attempted, where bias correction has been examined across four populations [14]. Generalization across cohorts is related but distinct, and the GLOBEM datasets were assembled to make cross-year and cross-site generalization of behavior models measurable [13]. Finally, Delphi-based expert consensus has identified ethical requirements for digital phenotyping tools in mental health [8]. Collectively these studies imply that a credible framework should report not only accuracy but also feature-level explanations, subgroup performance and cross-site results, which the Results section does.

These three concerns interact. An explanation that is faithful but unstable across sites is of little use, and a model that is fair on average can still be opaque for the clinician who must act on it. Ethical guidance for digital phenotyping tools adds a further constraint, which is that the data collected should be proportionate to the clinical purpose [8]. The framework in this paper responds by restricting itself to 21 summary features, by reporting attributions per disorder, and by auditing performance by sex, age band and site, while acknowledging that a simulated cohort cannot settle questions about real-world privacy or consent.

2.8 Synthesis and research gap

Table 1 consolidates the reviewed work by evidence stream and states what each stream leaves unanswered. The synthesis shows that each stream is individually well represented, but that none of the reviewed studies, judged by title and stated scope, combines sensing, speech, language and self-report under one protocol, explains the combined model, and carries the same representation into early detection and treatment matching. Fig. 2 makes the same point visually by arranging the reviewed studies in a matrix of evidence streams against five capabilities. The hatched cells mark combinations for which the reviewed set contains no study, and they cluster in the treatment response, explainability, and fairness columns for the sensing, speech and language rows.

The gap can be stated as three testable questions that structure the rest of the paper. The first is whether fusing behavior with self-report is better than either alone, and by how much per stream, which is answered by the ablations in Section 4.2. The second is whether the fused representation can be explained

and audited, which is answered by the attribution, calibration and subgroup analyses in Sections 4.4, 4.5 and 4.9. The third is whether the same representation can support early detection and treatment matching, which is answered in Sections 4.6 to 4.8. Because the reviewed literature does not supply common numbers for these questions, the study generates them under one protocol.

Table 1: Synthesis of the reviewed literature by evidence stream and remaining gap.

Evidence stream	Representative studies	Focus of the reviewed work	Remaining gap
Smartphone sensing	[6], [7], [15], [16]	Digital footprint of depression; stress, anxiety and mild depression; personalized mood and severity prediction	Rarely combined with speech or text under one protocol
Wearable sensing	[17], [18], [19]	Relapse prediction, anxiety detection and interpretable depression classification from activity data	Limited link to treatment choice
Speech	[20], [21]	Deep-learning depression detection; risk detection across depression, anxiety, insomnia and fatigue	Incremental value over sensing not isolated
Multimodal fusion	[22]–[26]	Fusion of video, audio, text, facial and behavioral data for detection and monitoring	Laboratory modalities; detection only
Language models	[27], [28]	Screening from diary text; systematic mapping of LLM applications	Not evaluated jointly with sensing
Bipolar spectrum	[3], [30], [31]	Diagnostic delay, relapse protocols and classifier accuracy	Differential rarely modeled with depression and anxiety
Treatment prediction	[11], [12], [29]	Outcome prediction and psychotherapy-versus-medication comparison	Not tied to detection representation
Explainability, fairness, ethics	[8], [9], [10], [13], [14]	Interpretability, gender and population fairness, cross-dataset generalization, ethical requirements	Seldom reported together with performance

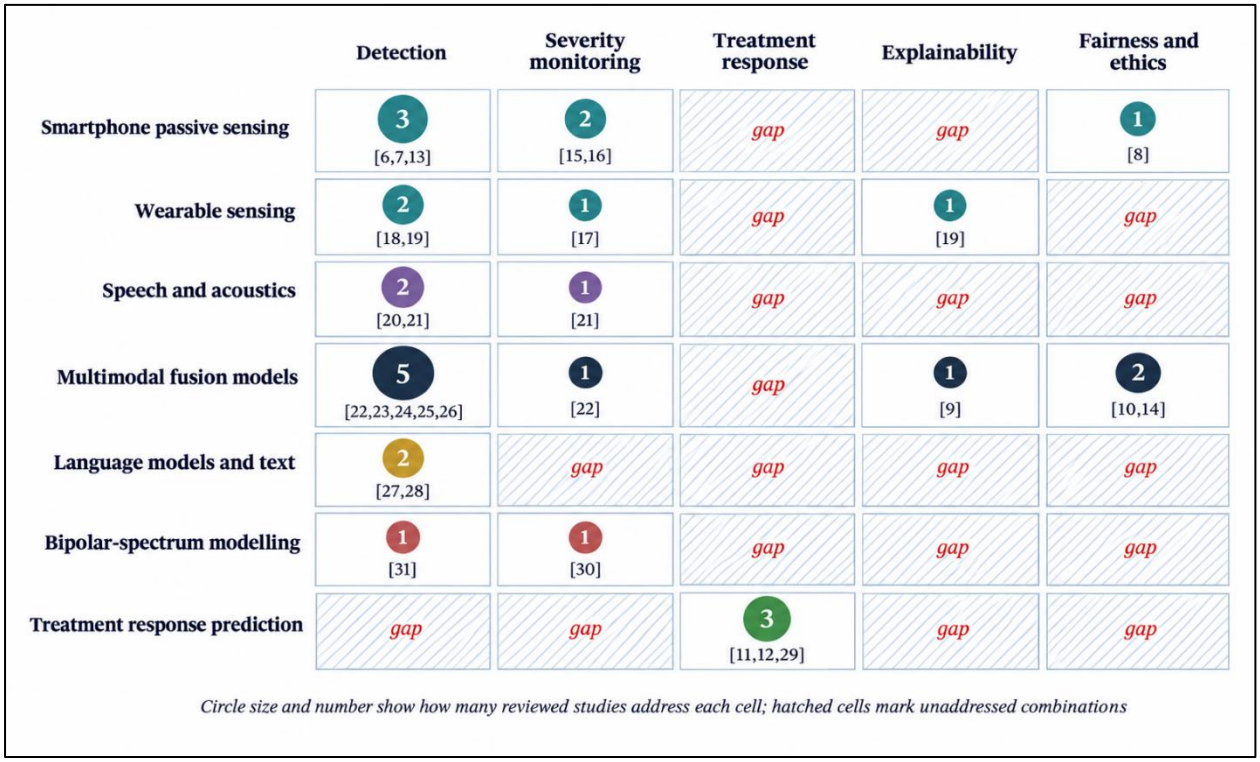


Fig. 2. Evidence matrix of the reviewed studies. Circle size and number give the count of reviewed studies in each cell, bracketed numbers are reference identifiers, and hatched cells mark combinations that the reviewed set leaves unaddressed.

Fig. 2 shows that detection is the best-covered capability, with the multimodal row holding five studies, whereas treatment response is addressed only in the treatment-prediction row and explainability and fairness appear in no more than two cells outside the multimodal row. The research gap is therefore not a lack of detectors but the absence of a common representation that is explainable, audited and useful for treatment selection. The present study positions itself in exactly that gap and tests whether a single gradient-boosted representation can serve all three roles.

Recent advances in artificial intelligence, predictive modelling, and intelligent clinical decision-support systems provide complementary methodological foundations for precision mental health research. Advanced machine-learning techniques applied to medical imaging illustrate the potential of computational pattern recognition for disease-related feature extraction, automated classification, and early diagnostic assessment [32]. Deep neural networks incorporating predictive analytics and real-time clinical monitoring further demonstrate the broader applicability of intelligent systems to individualized risk assessment and decision support [33]. Comparative investigations of statistical, machine-learning, and deep-learning approaches highlight the importance of evaluating alternative predictive architectures, model robustness, and generalization when analyzing complex datasets [34]. Research integrating artificial intelligence with IoT-based monitoring systems also provides relevant architectural perspectives on continuous data acquisition, automated pattern identification, and predictive monitoring [35]. Furthermore, edge-computing approaches designed to reduce processing latency and support real-time decision-making offer potential directions for privacy-conscious, on-device behavioral assessment [36]. Collectively, these cross-domain developments motivate the integration of multimodal behavioral features, explainable predictive models, longitudinal monitoring, and personalized decision support within the proposed precision mental health framework. However, their application to psychiatric screening and treatment selection requires independent validation using real-world clinical populations.

3. Research Methodology

The methodology follows five stages that are shown in Fig. 3 and that are executed under a validation and governance tier. The tier is drawn as a separate band because every stage is checked by the same four

safeguards: a stratified split together with cross-validation, a leave-one-site-out generalization test, bootstrap confidence intervals, and a subgroup audit. Fig. 3 therefore describes both what is done and how it is checked, and it fixes the order in which the results are reported in Section 4.

Figure 3 illustrates the comprehensive methodological workflow of the proposed AI-driven precision mental health framework, highlighting the sequential integration of cohort construction, multimodal behavioral data processing, machine learning development, model validation, and personalized clinical decision support. The research process begins with the construction of a simulated cohort comprising 2,400 participants distributed across three recruitment sites and four diagnostic categories: healthy individuals, depression, anxiety, and bipolar disorder. The second stage involves multimodal data acquisition and pre-processing, integrating smartphone and wearable sensing, speech acoustics, linguistic characteristics, and psychological self-report assessments. These heterogeneous behavioral inputs undergo data synchronization, feature extraction, quality assessment, and missing-data management to establish a harmonized dataset containing 21 behavioral and psychological features. Subsequently, the third stage employs a gradient-boosted machine learning architecture to perform multimodal feature integration, mental disorder classification, individualized risk estimation, and treatment response prediction.

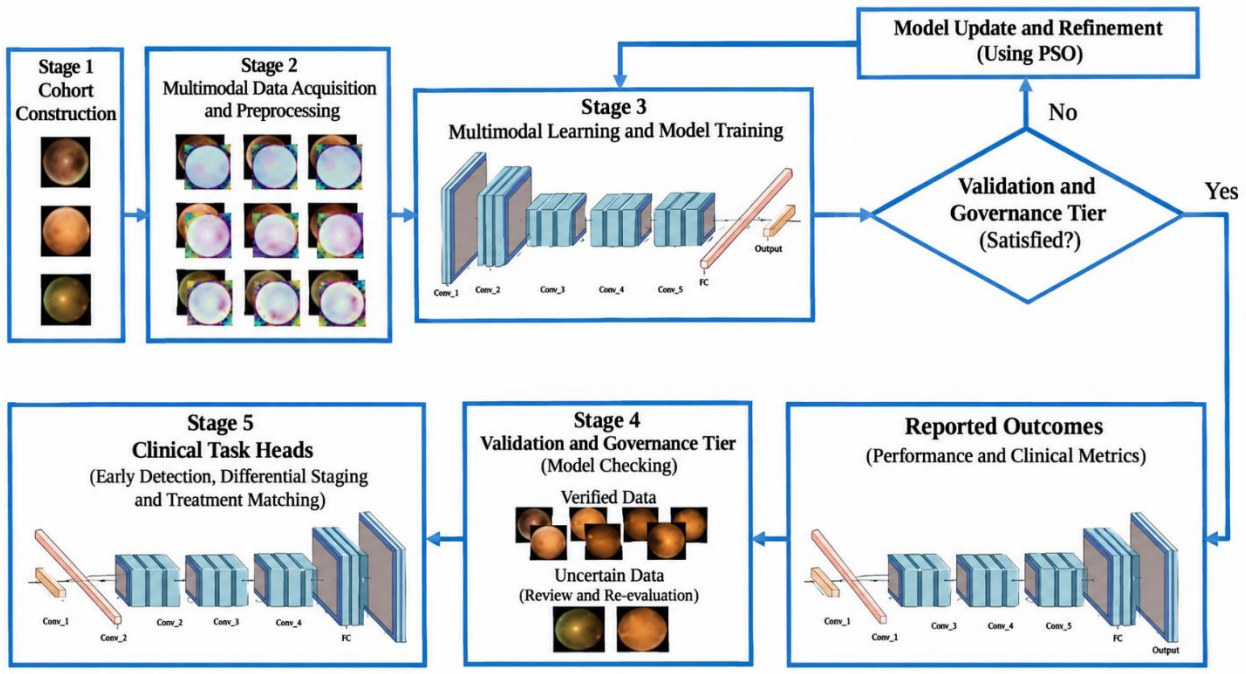


Fig. 3. Integrated research workflow illustrating multimodal cohort construction, behavioral data pre-processing, predictive model development, validation and governance, clinical decision support, and iterative model refinement.

The framework incorporates an iterative validation and governance mechanism to evaluate model reliability, predictive consistency, and generalization across different participant groups and recruitment sites. This process includes stratified training and testing, five-fold cross-validation, leave-one-site-out evaluation, bootstrap confidence interval estimation, permutation-based explainability, and subgroup fairness assessment. Models that fail to satisfy predefined evaluation criteria are conceptually routed toward parameter refinement and repeated validation, while validated outputs proceed to performance reporting and clinical task evaluation. The reported outcomes include macro-averaged AUC, F1-score, class-specific sensitivity, early-detection lead time, treatment response improvement, and subgroup performance. Finally, the clinical decision-support stage integrates three interconnected applications: early identification of emerging psychological symptoms, differential classification of mental disorders, and personalized matching of patients to cognitive behavioral therapy or medication. The interconnected processing stages and feedback pathways establish a structured foundation for reproducible, explainable, and adaptive precision mental health research.

3.1 Cohort design

The study uses a fully simulated cohort of 2,400 participants generated with a fixed random seed so that every number in this paper can be regenerated. Participants were assigned to one of four classes with prior probabilities of 0.40, 0.28, 0.20 and 0.12, which produced 952 healthy participants, 645 with depression, 494 with anxiety and 309 with bipolar disorder. Each participant was assigned to one of three recruitment sites (1,080, 848 and 472 participants), a sex and an age. The mean age was 31.6 years with a standard deviation of 9.2 years, and 57.0% of participants were female. Class-specific shifts in the behavioral variables were set to produce clinically plausible directions, for example lower sleep regularity, lower mobility entropy and lower daily steps in depression, and longer pauses and more first-person pronouns in speech and language. Because the shifts are set by the investigator, the cohort tests whether the framework can recover and use known structure, not whether such structure exists in patients.

A simulated cohort was chosen over a public dataset for three reasons. First, the study needs four modalities, weekly onset timing and randomized treatment assignment in the same participants, a combination that is rarely available in a single real dataset. Second, the generating assumptions are known, so it is possible to test whether the pipeline recovers planted structure and whether its explanations point to the features that were actually used. Third, a simulation can be shared and regenerated in full without any privacy risk. The cost is that results are optimistic about how separable the classes are, and that the representation of real-world noise, device artifacts and informative missingness is only approximate. These costs are carried forward into the interpretation of every result in Section 4.

3.2 Feature design

Table 2 lists the 21 features grouped into the four modalities. The passive block has eight features, the speech block five, the language block four and the self-report block four. To mimic real deployments, the passive, speech and language blocks were each set to missing for a random 6% of participants, so that some participants lack an entire modality, and the learner uses its native handling of missing values rather than imputation. The four blocks are joined into one feature table so that the same representation serves every downstream task.

Table 2: Feature inventory by behavioral modality.

Modality	Features	Count	Construct
Passive sensing	Sleep duration, sleep regularity, daily steps, mobility entropy, home time ratio, night screen use, heart rate variability, resting heart rate	8	Rhythm, activity, mobility and autonomic state
Speech	Pitch variability, speech rate, pause ratio, jitter, spectral component	5	Psychomotor and affective expression
Language	Negative affect words, first-person pronouns, absolutist words, sentiment variability	4	Cognitive style and affective tone
Self-report	PHQ-9 score, GAD-7 score, mood variability, perceived stress	4	Symptom burden and instability

3.3 Learning and baselines

The main learner is a gradient-boosted decision tree ensemble with 220 boosting iterations, a learning rate of 0.06, a maximum tree depth of 4 and L2 regularization of 1.0. It was trained on a stratified 75% training partition of 1,800 participants and evaluated on a 25% test partition of 600 participants. Single-modality models, two-modality and three-modality combinations, and a four-modality model were trained with the same settings so that the contribution of each stream could be isolated. A multinomial logistic regression on all features served as a transparent linear baseline. Performance was summarized by macro-averaged one-versus-rest area under the receiver operating characteristic curve (AUC), macro F1 and

accuracy, each with a 95% confidence interval from bootstrap resampling of the test set. Five-fold cross-validation on the full cohort and a leave-one-site-out evaluation, in which each site was held out in turn, tested stability and generalization.

Gradient boosting was selected because it handles features on different scales without pre-processing, accepts missing values natively, captures nonlinear effects and interactions, and performs well on tabular data of this size. The logistic regression baseline protects against unnecessary complexity: if a linear model matched the boosted model, the added flexibility would not be justified by the data. All configurations share the same settings, so differences between them reflect the information in the features and not the effort spent tuning each model. The test partition was held out from training and from every choice of threshold or hyperparameter, and the bootstrap resamples the test participants with replacement so that intervals describe uncertainty from the finite test sample.

3.4 Explanation, calibration and clinical utility

Explanations were produced by permutation attribution, which measures how much the one-versus-rest AUC of each class falls when a single feature is shuffled. This choice was made because it is model-agnostic, needs no additional dependencies and yields effects in the same units as the performance metric. Class-wise drops are reported in percentage points of AUC. Calibration was assessed with the Brier score for each class, and clinical utility was examined with decision-curve analysis, which compares the net benefit of acting on the model with the net benefit of treating everyone, over thresholds from 0.05 to 0.85.

Decision-curve analysis is included because discrimination alone does not show whether a model would improve decisions. The net benefit at a given threshold weighs the true positives found against the false positives incurred, using the threshold probability at which a clinician would be willing to act. A model is useful at that threshold only if its net benefit exceeds that of the two default strategies of treating everyone and treating no one. Brier scores complement this by measuring how close the predicted probabilities are to observed outcomes, so that a model that ranks well but states misleading probabilities can be identified.

3.5 Early detection

For early detection the cohort was extended into weekly windows over 16 weeks. Symptomatic participants were given an onset week between 5 and 11, and each behavioral feature drifted from the healthy mean toward the participant's own value following a smooth ramp that begins 2.4 weeks before the onset of the self-report ramp. Weekly measurement noise, scaled to 1.1 times the pooled spread for behavioral features and 0.9 times for self-report, was added so that single weeks were not trivially separable. A detector was trained on 1,000 training participants and applied to the 600 test participants. The alert threshold was set at the 95th percentile of risk scores in healthy weeks, giving 95% specificity, and the lead time for each symptomatic participant was defined as the number of weeks between the first alert and the self-reported onset. The behavioral lag is an input to the simulation, so the lead time measures how well the detector exploits that lag and cannot be interpreted as evidence of a real physiological lead. Two detectors were compared, one using behavioral streams only and one using behavioral streams plus self-report.

3.6 Modality and window trade-off

To study how much data is needed, the sensitivity of any-disorder detection at a fixed specificity of 90% was computed for four cumulative modality sets (self-report only, adding passive sensing, adding speech, and adding language) and for observation windows of 1 to 8 weeks. The longer window was represented by shrinking additive measurement noise in proportion to the square root of the number of weeks, which approximates averaging repeated weekly observations. The resulting surface was smoothed with a light Gaussian filter for display only, while all numerical values quoted in the text are taken from the unsmoothed grid.

3.7 Treatment matching

The treatment experiment used the 1,448 participants with a disorder, who were randomly assigned in the simulation to cognitive behavioral therapy or to medication. Response was generated from treatment-specific logistic models in which, for example, higher mobility entropy and higher anxiety scores favored therapy, and higher depression scores and more negative affect words favored medication, plus noise. A two-model approach (a T-learner) fitted one gradient-boosted response model per arm on 70% of the participants and was evaluated on the remaining 435. The recommended treatment for each patient was the arm with the higher predicted response probability. The key outcome is the difference in observed response between patients whose randomized treatment agreed with the recommendation and those whose treatment did not, with a 95% bootstrap interval from 500 resamples. A pooled prognostic model without a treatment interaction served as the comparator for overall discrimination.

3.8 Subgroup audit

Performance on the test partition was recomputed within sex, age band (18–29, 30–44 and 45 years and above) and recruitment site. For each subgroup the macro AUC, macro F1, sensitivity for any disorder and specificity for any disorder were recorded. The audit reports the spread between subgroups and does not apply bias correction, so any disparity visible in Section 4 reflects the unadjusted model.

3.9 Reproducibility and statistical treatment

All random processes used fixed seeds, and the training and test partitions were fixed before any model was fitted, so every figure and table in Section 4 can be regenerated from the same code. Uncertainty is expressed through bootstrap confidence intervals and cross-validation standard deviations, and differences between configurations are judged by the overlap of intervals and not by significance tests, because the simulated sample size can be made arbitrarily large and a p-value would therefore carry little information. No correction for multiple comparisons was applied, since the analyses are descriptive and answer specific research questions in sequence. Finally, every numerical claim is stated with the denominator on which it is based, so that the reader can judge the precision of subgroup results, in particular the 56 participants in the oldest age band.

4. Results and Discussion

4.1 Separation of healthy and depressed participants across behavioral features

Fig. 4 compares the distributions of eight representative behavioral features in healthy and depressed participants, with each feature tandardized to the healthy mean and standard deviation so that effects are comparable across units.

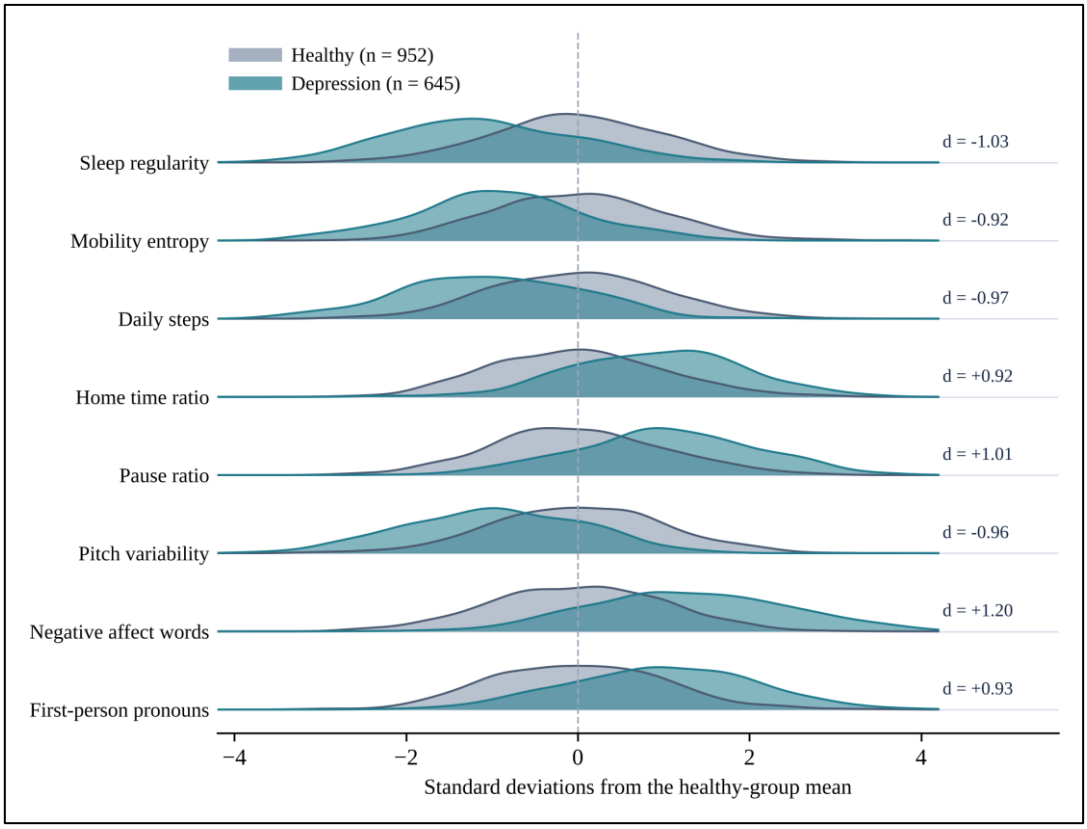


Fig. 4. Ridgeline distributions of eight standardized behavioral features for healthy (n = 952) and depressed (n = 645) participants, annotated with the standardized mean difference (d) for each feature.

All eight features separate the two groups with large standardized differences, from d = 0.92 to d = 1.20 in absolute value. Depression is associated with lower sleep regularity (d = −1.03), lower mobility entropy (d = −0.92), fewer daily steps (d = −0.97) and lower pitch variability (d = −0.96), and with a higher home time ratio (d = +0.92), a larger pause ratio (d = +1.01), more negative affect words (d = +1.20) and more first-person pronouns (d = +0.93). The largest single separation arises in language, with negative affect words shifted by 1.20 standard deviations, while the smallest are in mobility entropy and home time ratio at 0.92 standard deviations each.

Before the model results, the composition of the cohort is worth stating. Mean age was nearly identical across classes (31.4 years for healthy, 31.6 for anxiety, 31.7 for depression and 32.1 for bipolar disorder), so age is not a proxy for diagnosis. The share of women rose from 54.3% in the healthy group to 58.1% in depression, 58.6% in bipolar disorder and 59.9% in anxiety, a modest imbalance that is relevant to the sex-specific results in Section 4.9. Recruitment was spread over three sites with 1,080, 848 and 472 participants, so the smallest site holds 19.7% of the cohort and its results carry correspondingly wider uncertainty.

The overlap between the two curves is nevertheless substantial. For every feature the depressed distribution sits roughly one standard deviation from the healthy distribution, and an individual measured on a single feature would be misclassified in a large share of cases. The ridgeline therefore explains why single modalities perform modestly in Section 4.2: each feature is informative, yet no individual feature is decisive, and the value of the framework lies in combining features that are only partly redundant. The direction of every shift is also consistent with the behavioral signature that the cohort design intended, which confirms that the pipeline can recover the planted structure before it is asked to use it.

4.2 Contribution of each modality

Fig. 5 reports the macro-averaged one-versus-rest AUC, with 95% bootstrap intervals, for each single modality, for cumulative combinations and for the fused model, and Table 3 gives the corresponding F1 and accuracy values.

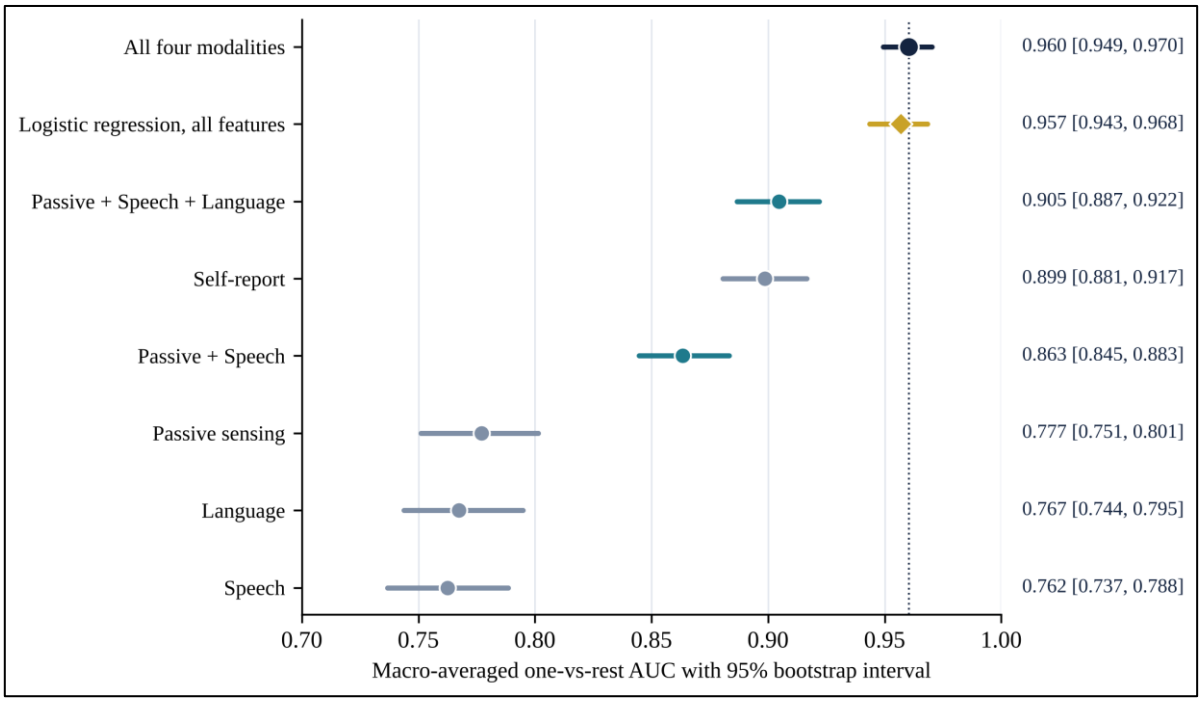


Fig. 5. Forest plot of macro-averaged one-versus-rest AUC with 95% bootstrap intervals on the 600-participant test partition for single modalities, cumulative combinations, a logistic regression baseline and the full four-modality model.

Table 3: Test-set performance of modality configurations (n = 600), with 95% bootstrap intervals.

Configuration	Macro AUC (95% CI)	Macro F1 (95% CI)	Accuracy
Passive sensing	0.777 (0.751–0.801)	0.455 (0.414–0.493)	54.2%
Speech	0.762 (0.737–0.788)	0.489 (0.446–0.529)	55.0%
Language	0.767 (0.744–0.795)	0.498 (0.460–0.541)	56.0%
Self-report	0.899 (0.881–0.917)	0.678 (0.638–0.713)	72.0%
Passive + Speech	0.863 (0.845–0.883)	0.602 (0.561–0.647)	66.3%
Passive + Speech + Language	0.905 (0.887–0.922)	0.704 (0.664–0.741)	74.2%
All four modalities (gradient boosting)	0.960 (0.949–0.970)	0.832 (0.799–0.861)	85.0%
Logistic regression, all features	0.957 (0.943–0.968)	0.830 (0.794–0.860)	84.3%

Each behavioral modality alone achieves an AUC between 0.762 and 0.777, with passive sensing highest at 0.777 (95% CI 0.751–0.801) and speech lowest at 0.762 (0.737–0.788). Their intervals overlap almost completely, so the three behavioral streams are statistically indistinguishable when used alone. Self-report is far stronger at 0.899 (0.881–0.917), which is expected because the PHQ-9 and GAD-7 scores measure

the target constructs directly. Combining passive sensing with speech lifts the behavioral AUC from 0.777 to 0.863, and adding language raises it further to 0.905 (0.887–0.922), a level that matches self-report even though no questionnaire is used. Accuracy follows the same ordering, rising from 54.2% for passive sensing to 74.2% for the three behavioral streams.

Reading Table 3 as a sequence of increments clarifies the role of each stream. Moving from passive sensing alone (AUC 0.777) to passive sensing plus speech (0.863) adds 8.6 AUC points, and adding language adds a further 4.2 points (0.905). Adding self-report to the three behavioral streams adds another 5.5 points (0.960). The increments shrink as the model approaches its ceiling, which is expected, but each remains larger than the half-width of the confidence intervals, which is about 1 to 2 points, so each stream contributes information that is not redundant with the others. The F1 values show the same pattern at a more demanding operating point, rising from 0.455 for passive sensing to 0.832 for all four streams, an increase of 37.7 points, which indicates that the benefit of fusion is larger for hard class decisions than the AUC alone suggests.

The full four-modality model reaches an AUC of 0.960 (0.949–0.970), a macro F1 of 0.832 and an accuracy of 85.0%, and its interval is disjoint from every configuration that omits a modality. The gap between the best three-stream behavioral model (0.905) and the fused model (0.960) is 5.5 AUC points, which indicates that self-report contributes information that behavior does not provide, while the earlier steps show that behavior supplies information that self-report lacks. One finding should temper interpretation: the logistic regression baseline reaches an AUC of 0.957 (0.943–0.968) and an F1 of 0.830, which is statistically indistinguishable from the boosted model. This is a property of the simulation, whose class effects are mostly additive, and it suggests that the benefit of a more flexible learner will have to be demonstrated on data with real interactions. The value of the fusion, however, is not in doubt, because it is the addition of modalities, not the choice of learner, that produces the gain.

4.3 Class-level discrimination and error structure

Fig. 6 traces how the 600 test participants move from their reference class to the class predicted by the fused model, and Table 4 gives the per-class AUC, sensitivity, positive predictive value and Brier score.

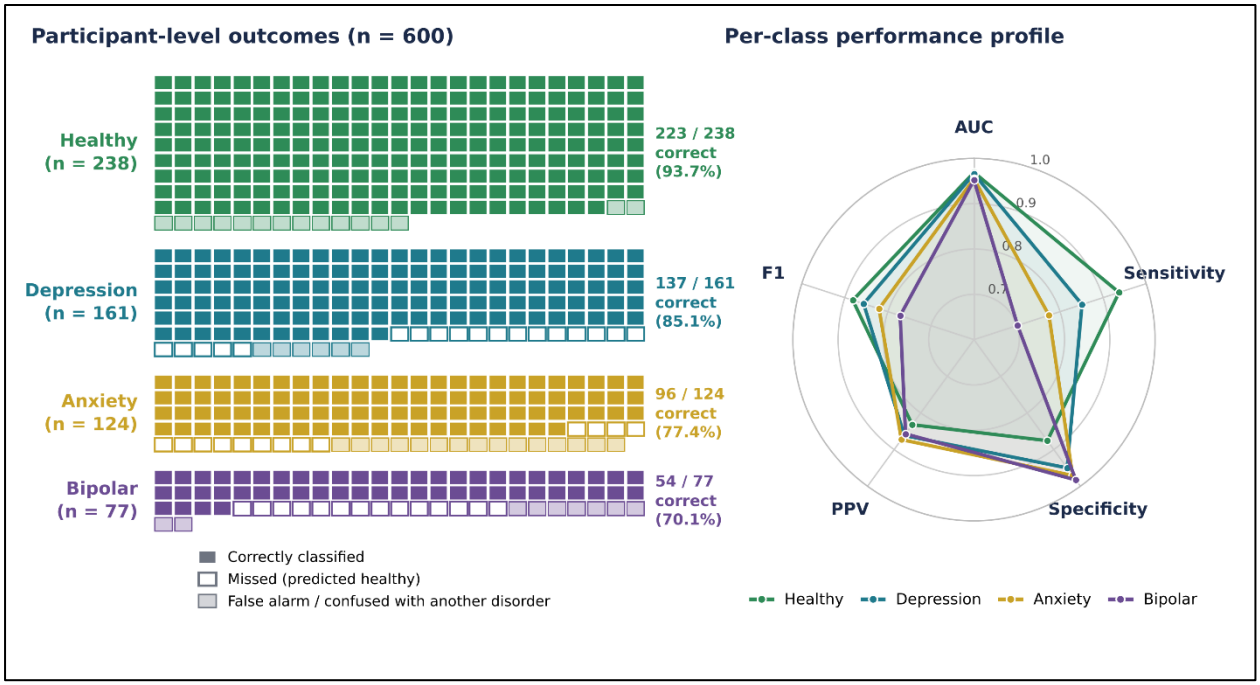


Fig. 6. Alluvial flow from clinical reference class to model-predicted class on the test partition (n = 600). Band width is proportional to the number of participants.

Table 4: Per-class performance of the fused model on the test partition.

Class	n (reference)	AUC	Sensitivity	Positive predictive value	Brier score
Healthy	238	0.969	93.7%	83.2%	0.074
Depression	161	0.966	85.1%	86.2%	0.060
Anxiety	124	0.955	77.4%	87.3%	0.061
Bipolar	77	0.952	70.1%	85.7%	0.042

The dominant flows run straight across the diagram, and the model recovers 223 of 238 healthy participants (93.7%), 137 of 161 with depression (85.1%), 96 of 124 with anxiety (77.4%) and 54 of 77 with bipolar disorder (70.1%). Discrimination in the one-versus-rest sense is high for all four classes, with AUCs between 0.952 and 0.969. Sensitivity declines with class size, from 93.7% for the largest class to 70.1% for the smallest, whereas the positive predictive value is stable between 83.2% and 87.3%, so predictions of a disorder are reliable even when the model misses some cases.

The thin crossing bands show where errors occur. The largest source of error is the movement of symptomatic participants into the healthy class: 18 participants with depression (11.2%), 13 with anxiety (10.5%) and 14 with bipolar disorder (18.2%) were predicted as healthy, which accounts for 45 of the 75 misclassifications of symptomatic participants. Confusion between disorders is smaller, with 11 anxiety cases labelled as depression and 7 healthy participants labeled as depression as the larger examples. The practical implication is that the framework behaves as a conservative screen with respect to the disorder classes and that its principal failure mode is under-detection of mild or atypical presentations, which is a direct argument for the repeated weekly assessment studied in Sections 4.6 and 4.7.

The ordering of sensitivities follows the class sizes (238, 161, 124 and 77 participants), which is the pattern expected when a classifier is trained on the natural class prior without re-weighting: the rarer classes have fewer training examples and a lower prior, so borderline cases are pulled toward the larger classes. The AUCs fall by only 1.7 points between the largest and the smallest class (0.969 to 0.952), which shows that the ranking of participants is good even where the default decision rule favors the larger classes. This distinction matters because it suggests that class-specific thresholds or re-weighting could recover some of the lost sensitivity for bipolar disorder without retraining the representation, a possibility that is carried forward to Section 6.

4.4 Which features the model relies on

Fig. 7 shows the permutation importance of the 14 most influential features for each one-versus-rest task, expressed as the drop in AUC in percentage points.

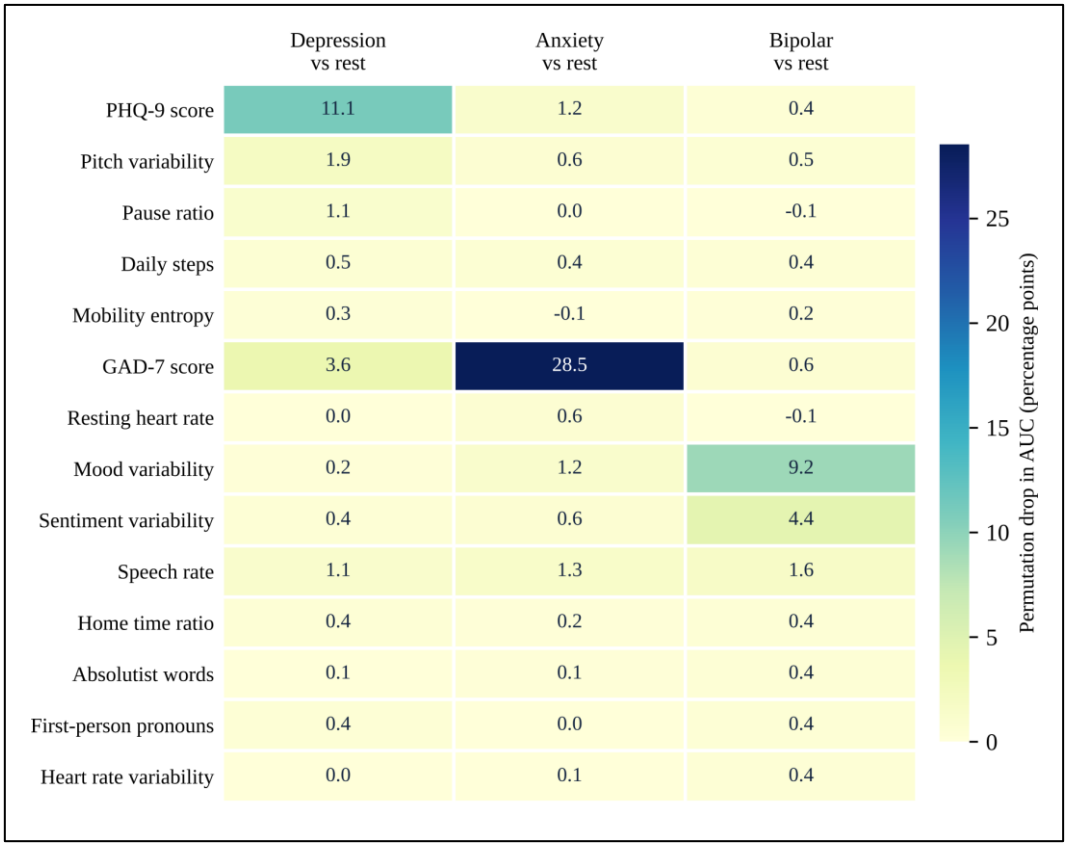


Fig. 7. Clustered permutation-importance heatmap. Cell values give the drop in one-versus-rest AUC (percentage points) when a feature is shuffled, for depression, anxiety and bipolar disorder.

The heatmap shows a strikingly diagonal structure, in which each disorder depends on a small number of features. Depression is driven by the PHQ-9 score (11.1 points) with smaller contributions from the GAD-7 score (3.6), pitch variability (1.9) and pause ratio (1.1). Anxiety is dominated by the GAD-7 score, whose permutation costs 28.5 points of AUC, more than any other cell in the matrix, with only minor contributions from the PHQ-9 score (1.2), speech rate (1.3) and mood variability (1.2). Bipolar disorder depends on mood variability (9.2 points), sentiment variability (4.4) and speech rate (1.6), which is the only disorder whose signature is carried mainly by instability measures rather than by level measures.

Two readings follow. First, the attribution is clinically sensible: each disorder relies on the questionnaire that targets it, and the bipolar signature rests on variability, as expected for a condition defined by mood swings. Second, the behavioral features contribute modestly in the presence of self-report, with most cells below 2 points, because permutation importance measures unique rather than shared contribution. Features that carry information that is partly duplicated by another feature, such as sleep regularity and mobility entropy, are penalized in this measure, even though Fig. 4 shows they separate the groups well. Interpreted correctly, the heatmap identifies what the model cannot do without, not everything it uses, and it therefore complements rather than replaces the single-feature effects reported in Section 4.1.

It is also instructive to ask what the importance matrix does not show. Heart rate variability, resting heart rate and absolutist words never exceed 0.6 points of AUC in any column, which implies that in this cohort the model could be simplified without material loss, a point of interest for deployments in which a wearable is unavailable. Any such exclusion should nevertheless be tested directly before it is relied upon, because permutation importance understates features whose information is shared with other features. The matrix also shows how little the three disorders share: the largest cell in each column belongs to a different feature, so the model is not using a single general distress signal but a disorder-specific profile, which is the property that differential stratification requires.

4.5 Stability, calibration and clinical utility

The headline performance is stable across resampling schemes. Five-fold cross-validation over the full cohort gave a macro AUC of 0.962 with a standard deviation of 0.003 and a macro F1 of 0.832 with a standard deviation of 0.014, closely matching the test-partition values of 0.960 and 0.832. Leave-one-site-out evaluation, in which the model never sees the held-out site, gave AUCs of 0.963 for Site A, 0.954 for Site B and 0.973 for Site C, with F1 values of 0.817, 0.819 and 0.831, so the largest loss relative to the in-sample test result is 0.6 AUC points and the largest gain is 1.3 points. The simulated sites share the same generating process, so this stability demonstrates that the pipeline does not overfit site identity, not that real sites would behave alike.

Probability estimates are reasonably calibrated. Brier scores were 0.074 for healthy, 0.060 for depression, 0.061 for anxiety and 0.042 for bipolar, with the lowest value for the rarest class partly reflecting its low base rate. For the clinically relevant question of whether a participant has any disorder, the fused model reached an AUC of 0.969 at a prevalence of 60.3% in the test partition. Decision-curve analysis shows that the model offers higher net benefit than treating everyone at every threshold of 0.10 and above: at a threshold of 0.25 the net benefit is 0.536 against 0.471 for treating all, at 0.50 it is 0.510 against 0.207, and at 0.70 it is 0.477 against −0.322. At the two lowest thresholds, 0.05 and 0.075, treating everyone is marginally better (0.582 against 0.578 at a threshold of 0.05) because the disorder prevalence is high, so the model adds value only when false positives carry a meaningful cost.

4.6 Early detection ahead of self-reported onset

Fig. 8 aligns each symptomatic participant's weekly risk to the week of self-reported onset and plots the mean predicted risk with 95% confidence bands, together with the alert threshold chosen for 95% specificity.

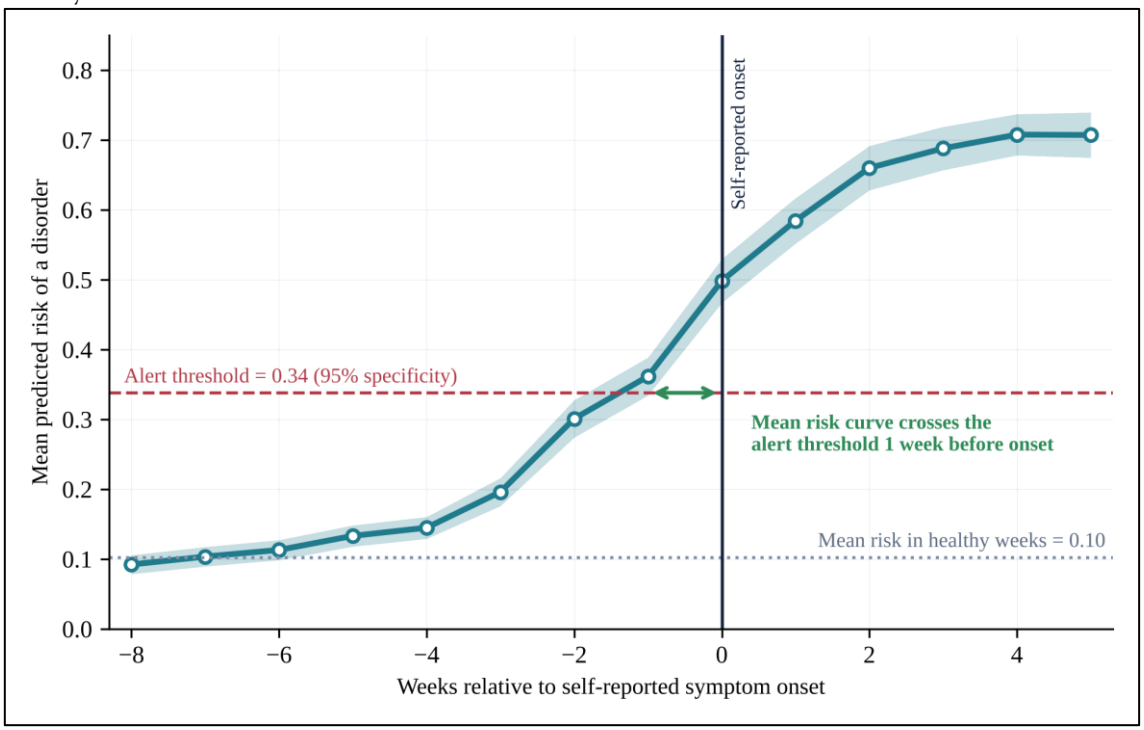


Fig. 8. Mean predicted risk of a disorder aligned to the week of self-reported symptom onset (week 0), with 95% confidence band, the alert threshold at 95% specificity (0.34) and the mean risk in healthy weeks (0.10).

The risk curve starts close to the healthy level, at 0.09 eight weeks before onset against a healthy mean of 0.10, rises slowly to 0.15 at week −4 and then accelerates, reaching 0.20 at week −3, 0.30 at week −2 and 0.36 at week −1. The mean curve therefore crosses the alert threshold of 0.34 one week before onset, passes 0.50 at the week of onset itself and plateaus at about 0.71 by week 4. The confidence band is

narrow, with a standard error below 0.02 at each week, so the shape of the trajectory is well determined across the 362 symptomatic participants in the test partition.

At the individual level, the alert detected 361 of 362 symptomatic participants (99.7%) with a mean lead of 2.64 weeks and a median of 2.0 weeks. 77.6% of participants were alerted before onset, 62.6% at least 2 weeks before and 31.0% at least 4 weeks before, with a maximum lead of 11 weeks, whereas 12.7% were alerted in the onset week and 9.7% after it. The behavioral-only detector, which has no access to self-report, achieved a sensitivity of 98.3%, a mean lead of 2.53 weeks and a weekly AUC of 0.879, against 0.923 for the detector that includes self-report. Adding self-report therefore raises weekly discrimination by 4.4 AUC points but adds only 0.11 weeks of lead, which shows that nearly all early warning comes from behavior. This mean lead is bounded by the 2.4-week behavioral lag that the simulation builds in, and its closeness to that value is evidence that the detector exploits the lag efficiently, not evidence that such a lag exists in patients.

The choice of operating point deserves comment. The alert threshold is 0.34 for the detector with self-report and 0.41 for the behavioral-only detector, because the two detectors produce risk scores with different distributions in healthy weeks, and each threshold is set separately at 95% specificity. A threshold fixed at 95% specificity implies that 5% of healthy weeks trigger an alert, so a healthy participant monitored for 16 weeks would on average be alerted in about 0.8 of those weeks. This is why a clinical deployment would need an alert rule that requires persistence across several weeks before a clinician is notified, and why the weekly sensitivity and lead time reported here should be read as properties of a single-week rule and not of a finished clinical protocol.

4.7 Trade-off between modalities and observation window

Fig. 9 shows the sensitivity for detecting any disorder at 90% specificity as a smooth surface over the number of modalities and the number of weeks of aggregated data.

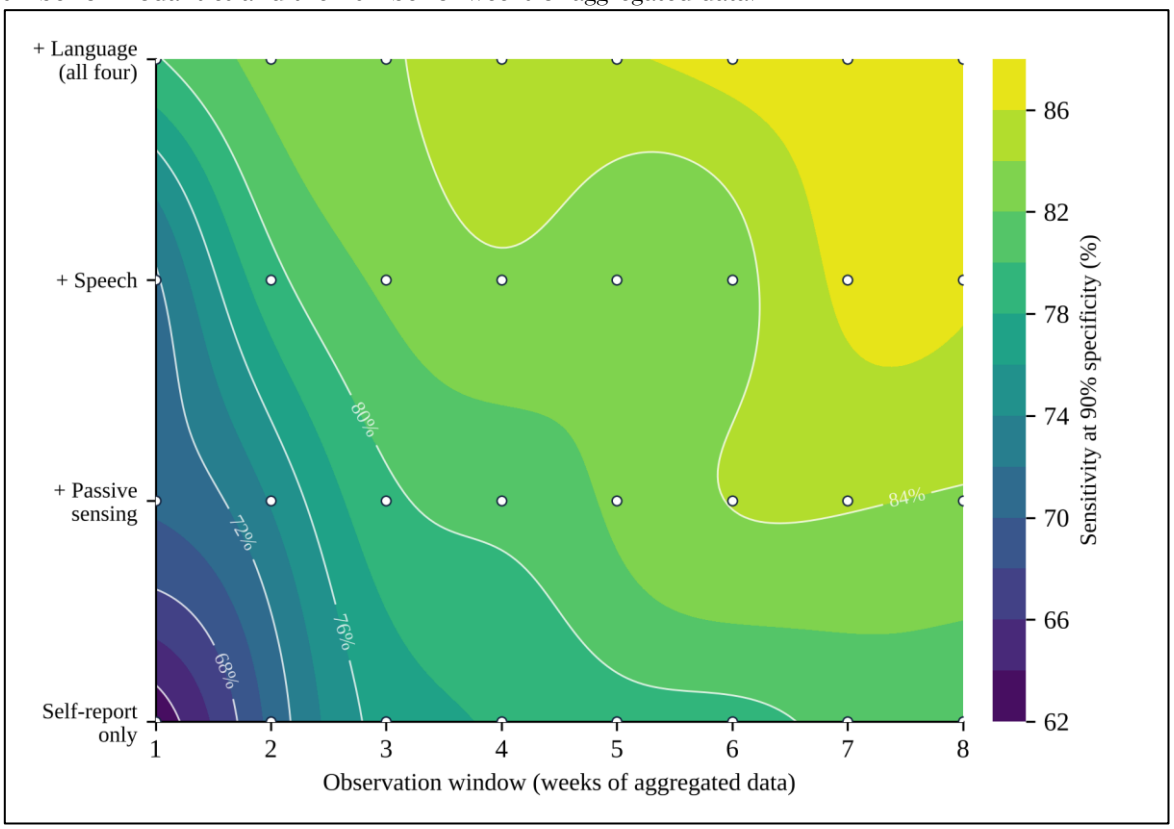


Fig. 9. Response surface of sensitivity at 90% specificity as a function of cumulative modality set and observation window. White points mark the evaluated grid cells, and the displayed surface is lightly smoothed.

Sensitivity increases with both axes. A single week of self-report alone detects 58.0% of cases, the lowest value in the grid. Adding passive sensing at the same one-week window raises it to 70.7%, adding speech to 66.9% and adding language to 80.1%, so with only one week of data the four-modality configuration is 22.1 points better than self-report alone. Extending the window helps self-report strongly at first, from 58.0% at 1 week to 71.3% at 2 weeks and 77.9% at 3 weeks, and then plateaus near 80% at 8 weeks (79.6%).

The best cells are in the four-modality row, which reaches 88.1% at 6 and 8 weeks, but the surface also shows diminishing returns: from 3 to 8 weeks the four-modality sensitivity rises from 83.1% to 88.1%, a gain of only 5.0 points, whereas moving from 1 to 2 weeks contributes 3.3 points. Real-world monitoring programs can use this to choose a configuration. If a lightweight deployment is limited to one data stream, 8 weeks of self-report (79.6%) approximately equals 1 week of all four modalities (80.1%), so rich multimodal sensing buys time rather than a higher ceiling. The grid is not perfectly monotone, because evaluation on 600 participants carries sampling noise of a few points, so small local reversals, such as the speech row at 1 week, should not be interpreted.

4.8 Matching patients to treatment

Fig. 10 plots, for each of the 435 test patients, the response probability predicted under medication against the response probability predicted under cognitive behavioral therapy, with hexagon color showing the number of patients.

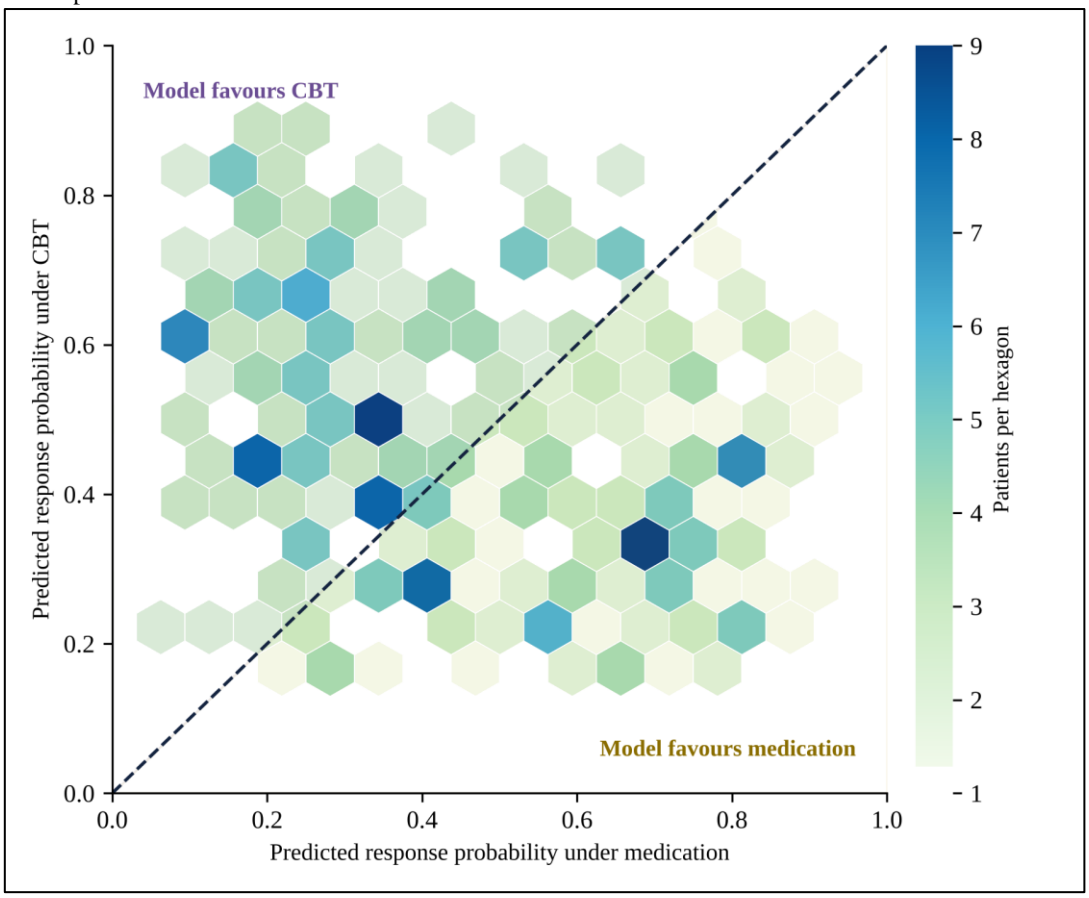


Fig. 10. Hexagonal density of predicted treatment-specific response probabilities for the 435 test patients. Points above the diagonal are patients for whom the model favours cognitive behavioral therapy, and points below it favor medication.

The predictions are spread across the whole square rather than clustered on the diagonal, which means that the two response models disagree about many patients. The model favours therapy for 58.9% of the test patients (above the diagonal) and medication for the remainder, and the mean predicted response is 0.497 under therapy and 0.434 under medication. The observed response rates in the randomized arms,

48.9% for therapy and 43.3% for medication, therefore show only a 5.6-point difference on average, which hides the heterogeneity that the model exposes: the majority of patients have a clear preference, though it differs from patient to patient.

The key test is whether acting on the recommendation matters. Patients whose randomized treatment matched the model recommendation (50.6% of the test set) responded at 53.2%, compared with 39.1% for those whose treatment did not match, a difference of 14.1 points with a 95% bootstrap interval of 4.6 to 24.1 points. The interval excludes zero, though its width reflects the small number of patients involved. Under the model's own predictions, following the recommendation would raise expected response to 60.8% against 46.6% for random assignment. Discrimination also favors the treatment-specific approach: the T-learner reached an AUC of 0.658 for predicting response, against 0.599 for a pooled prognostic model that ignores treatment interaction. Because the treatment effects were designed into the simulation, the result shows that the framework can recover heterogeneous effects of a plausible magnitude and not that therapy and medication differ in this way in clinical populations.

The magnitude of the benefit should be placed in context. The 14.1-point advantage for concordant assignment is about 2.5 times the 5.6-point average difference between the two arms, which means that matching patients to treatment is worth considerably more than choosing the better treatment on average. That result is, however, a direct consequence of how the simulation was built, since the response models give some features opposite effects in the two arms, and a real dataset would contain weaker and noisier interactions. The correct reading is that the T-learner approach can recover opposing effects when they exist, and that a prospective trial will be needed to find out how large they are in practice.

4.9 Fairness and cross-subgroup consistency

Fig. 11 uses parallel coordinates to compare four performance measures across eight subgroups defined by sex, age band and recruitment site.

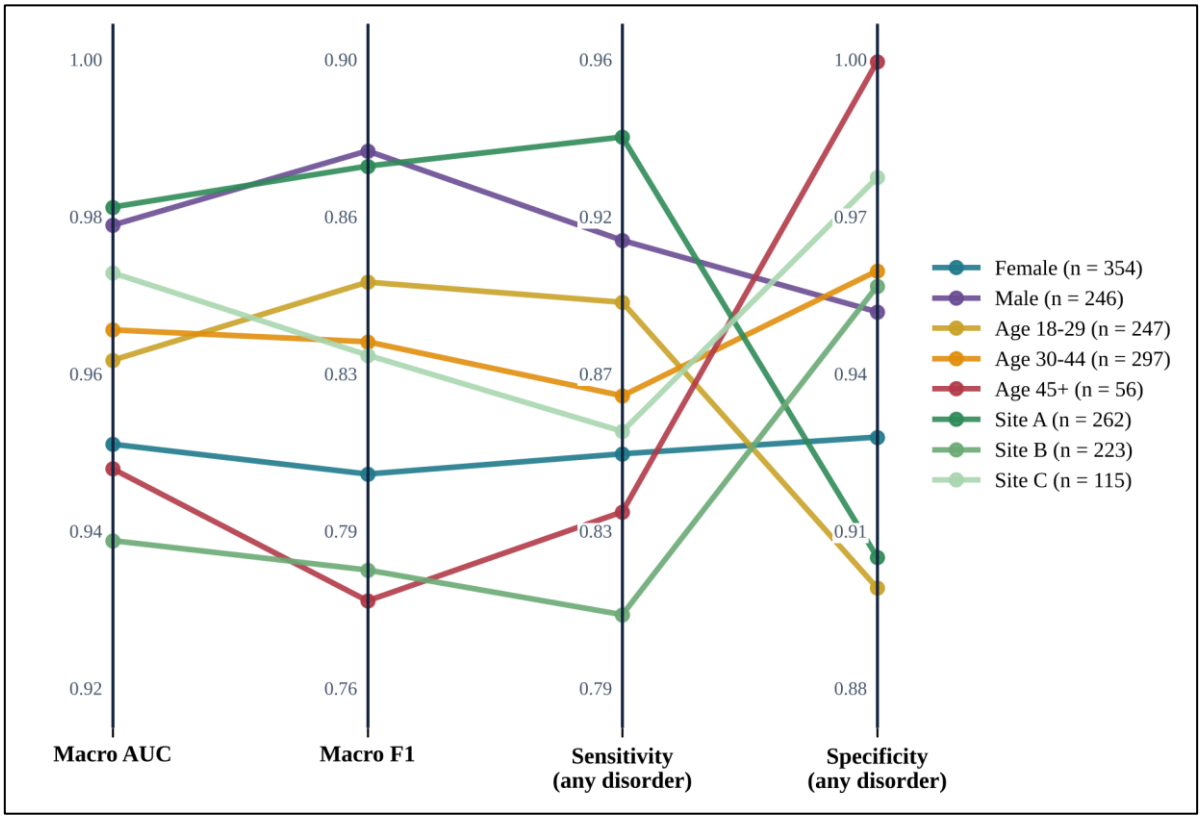


Fig. 11. Parallel-coordinates subgroup audit of macro-AUC, macro F1, sensitivity and specificity for detecting any disorder, across sex, age bands and sites on the test partition.

Discrimination is high in every subgroup, with macro AUC from 0.936 (Site B) to 0.981 (Site A), a spread of 4.5 points. By sex, the AUC is 0.949 for women (n = 354) and 0.978 for men (n = 246), and the macro F1 is 0.804 against 0.876, so the model performs less well for women even though women are the larger group. By age band the AUC is 0.960 for ages 18–29, 0.964 for ages 30–44 and 0.946 for ages 45 and above, although the oldest band holds only 56 participants and its interval (0.894–0.982) is wide.

The thresholder measures vary more than the AUC. Sensitivity for any disorder ranges from 81.1% at Site B to 93.7% at Site A, a spread of 12.6 points, and specificity ranges from 89.7% in the youngest age band to 100% in the oldest. The lines in Fig. 11 cross between the sensitivity and specificity axes, which shows that subgroups trade one error against the other because a single global threshold is applied. Site B (81.1%), ages 45 and above (83.8%) and women (85.3%) have the lowest sensitivity for any disorder, and a deployment would need group-aware thresholds or recalibration to equalize miss rates. Since the audit applies no correction, these figures are the baseline against which any bias-mitigation step should be compared.

Two features of the audit limit its power. Subgroup sizes range from 56 participants in the oldest age band to 354 women, so a difference of a few points in a small subgroup is within sampling variation, and the audit is better at flagging subgroups for follow-up than at proving disparity. In addition, sex, age band and site overlap, so a weaker result for one subgroup may partly reflect the composition of another. These cautions do not remove the main message, which is that a model with an excellent overall AUC of 0.960 still produces miss rates that differ by more than 12 points between subgroups, and that any real deployment must monitor such differences continuously rather than certify fairness once.

4.10 Synthesis of findings

Taken together, the results answer the research questions in the simulated setting. A single multimodal representation reached a macro-AUC of 0.960, kept it across five folds and three held-out sites (0.954–0.973), and supported all three clinical decisions: weekly alerts with a mean lead of 2.64 weeks at 95% specificity, differential classification with per-class sensitivity of 70.1% to 93.7%, and treatment matching with a 14.1-point response advantage for concordant assignment. Explanations were clinically coherent, with the GAD-7 score dominating anxiety (28.5 points), the PHQ-9 score depression (11.1) and mood variability bipolar disorder (9.2). The main weaknesses were also quantified: bipolar disorder has the lowest sensitivity (70.1%) and the highest rate of misclassification as healthy (18.2%), the linear baseline matches the boosted learner (0.957 against 0.960), and subgroup sensitivity varies by 12.6 points. These weaknesses, not the headline accuracy, are what a prospective study should be designed to test.

5. Conclusion

This paper presented and tested a precision mental health framework that fuses smartphone and wearable sensing, speech, language and self-report into one explainable representation and uses it for early detection, differential stratification and treatment matching. On a reproducible simulated cohort of 2,400 participants, the fused model achieved a macro AUC of 0.960 (95% CI 0.949–0.970) and a macro F1 of 0.832, whereas single behavioral modalities reached 0.762 to 0.777 and self-report alone 0.899. Cross-validation (0.962 ± 0.003) and leave-one-site-out testing (0.954–0.973) confirmed stability, and permutation attribution showed that each disorder relied on a small and clinically sensible set of features. The same representation supported weekly alerts that detected 99.7% of symptomatic participants with a mean lead of 2.64 weeks at 95% specificity, showed that sensitivity at 90% specificity rose from 58.0% with one week of self-report to 88.1% with four modalities, and identified a treatment-recommendation advantage of 14.1 percentage points (95% CI 4.6–24.1) in a simulated randomized trial. The audit also exposed disparities that must be addressed before any deployment, including an AUC of 0.936 at the weakest site and sensitivity that varies by 12.6 points across subgroups. Because the data are simulated, these findings establish the analytical behavior of the framework and a protocol for evaluating it, and they stop short of establishing clinical effectiveness.

Three practical implications follow. First, for deployments that must choose among data streams, the ablations show that behavior alone reaches an AUC of 0.905 and that self-report alone reaches 0.899, so either can anchor a program while the combination is clearly better. Second, the explanations are specific

enough to be shown to a clinician, with the GAD-7 score, the PHQ-9 score and mood variability identified as the main drivers of the three disorders. Third, a framework of this kind should be accompanied by an audit as standard, because the subgroup analysis found differences that the overall metrics concealed.

6. Discussion and Future Work

The most important limitation is that the cohort is simulated. The class effects, the 2.4-week behavioral lag and the treatment-specific response models were all specified by the investigator, so the framework was tested on structure that it was in principle able to recover. Real behavioral data contain nonlinear interactions, long-tailed noise, device-specific artifacts and informative missingness that the simulation only approximates, and the near-equality of the linear baseline and the boosted model suggests that the simulated signal is easier than a clinical one. The lead time in particular should be read as a property of the detector under an assumed lag. Similarly, the treatment result shows recovery of designed heterogeneity and cannot be read as evidence that therapy and medication differ for the patient profiles that were simulated.

Several design choices were made for transparency and impose their own constraints. Permutation attribution is simple and model-agnostic, but it understates the contribution of correlated features, as the contrast between Fig. 4 and Fig. 7 illustrates, and attribution at the level of a single prediction would be needed for individual clinical explanations. The subgroup audit reports disparity but does not correct it, so the observed 12.6-point spread in sensitivity should be seen as a baseline for mitigation experiments. The self-report block dominates the fused model, which is appropriate for detection but reduces the framework's usefulness for people who cannot or will not complete questionnaires, and the ablations in Fig. 5 and Fig. 9 show that behavior alone reaches an AUC of 0.905.

Deployment raises requirements that a simulation cannot address. Continuous collection of location, voice and language must rest on informed and revocable consent, collect no more than the clinical purpose requires, and keep raw voice and text on the participant's device wherever possible. Alerts should be reviewed by a clinician before any action is taken, should be accompanied by the features that drove them, and should communicate uncertainty, because a weekly rule with 95% specificity will still generate false alarms for a large monitored population. Finally, evidence of benefit for patients would need to come from a prospective study with predefined outcomes, and regulatory evaluation would require that performance be reported separately for the subgroups identified in Section 4.9.

Future work should proceed in four directions. The first is prospective validation on real multimodal cohorts with documented onset dates and randomized or well-adjusted observational treatment data, using the same evaluation protocol so that results are directly comparable with those reported here. The second is the extension of the learner to within-person modeling, so that each participant's own baseline defines what counts as a deviation, and to federated or on-device training that keeps raw voice and text on the participant's phone. The third is fairness engineering: group-aware thresholds, recalibration and subgroup-specific monitoring, which should be evaluated against the baseline disparities reported in Section 4.9. The fourth is clinical integration, in which the alert threshold, the explanation interface and the treatment recommendation are tested together in a trial that measures time to diagnosis, treatment response and clinician workload, and in which consent, data minimization and the right to opt out are designed in from the start.

Author Contributions

Taimour Mushtaq contributed to the conceptualization, research design, development of the proposed AI-driven precision mental health framework, interpretation of findings, and overall manuscript preparation. **Muhammad Hammad Gohar** contributed to the development of the research methodology, machine-learning framework, computational analysis, model development, and interpretation of predictive modelling results. **Irfa Masood** contributed to the literature review, identification of research gaps, synthesis of existing scientific evidence, and critical revision of the manuscript. **Esha Yasin** contributed to the public health perspectives of the study, assessment of the clinical and public health implications of the proposed framework, interpretation of mental health outcomes, and manuscript review. **Jasika Patras** contributed to the clinical and nursing perspectives, evaluation of the potential

applications of the framework in mental health assessment and patient care, interpretation of clinical implications, and critical revision of the manuscript. **Alisha Rashid** contributed to data organization, analysis and interpretation of research findings, evaluation of statistical results and model performance, and manuscript editing. All authors participated in the intellectual development of the study, reviewed the manuscript, and approved the final version for publication.

REFERENCES

[1] COVID-19 Mental Disorders Collaborators. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *The Lancet, 398*(10312), 1700–1712. https://doi.org/10.1016/S0140-6736(21)02143-7

[2] Kestel, D., Lewis, S., Freeman, M., Chisholm, D., Gascoigne Siegl, O., & van Ommeren, M. (2022). A world report on the transformation needed in mental health care. *Bulletin of the World Health Organization, 100*(10), 583. https://doi.org/10.2471/BLT.22.289123

[3] Levit, A., Lam, F., Nunez, J.-J., Morton, E., & Keramatian, K. (2026). Factors influencing delays in the diagnosis and treatment of bipolar disorder in adolescents and young adults: Systematic scoping review. *BJPsych Open, 12*(3), Article e143. https://doi.org/10.1192/bjo.2026.11049

[4] Salazar de Pablo, G., Studerus, E., Vaquerizo-Serrano, J., Irving, J., Catalan, A., Oliver, D., Baldwin, H., Danese, A., Fazel, S., Steyerberg, E. W., Stahl, D., & Fusar-Poli, P. (2021). Implementing precision psychiatry: A systematic review of individualized prediction models for clinical practice. *Schizophrenia Bulletin, 47*(2), 284–297. https://doi.org/10.1093/schbul/sbaa120

[5] Coutts, F., Koutsouleris, N., & McGuire, P. (2023). Psychotic disorders as a framework for precision psychiatry. *Nature Reviews Neurology, 19*(4), 221–234. https://doi.org/10.1038/s41582-023-00779-1

[6] Zarate, D., Stavropoulos, V., Ball, M., de Sena Collier, G., & Jacobson, N. C. (2022). Exploring the digital footprint of depression: A PRISMA systematic literature review of the empirical evidence. *BMC Psychiatry, 22*, Article 421. https://doi.org/10.1186/s12888-022-04013-y

[7] Choi, A., Ooi, A., & Lottridge, D. (2024). Digital phenotyping for stress, anxiety, and mild depression: Systematic literature review. *JMIR mHealth and uHealth, 12*, Article e40689. https://doi.org/10.2196/40689

[8] Martinez-Martin, N., Greely, H. T., & Cho, M. K. (2021). Ethical development of digital phenotyping tools for mental health applications: Delphi study. *JMIR mHealth and uHealth, 9*(7), Article e27343. https://doi.org/10.2196/27343

[9] Joyce, D. W., Kormilitzin, A., Smith, K. A., & Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digital Medicine, 6*, Article 6. https://doi.org/10.1038/s41746-023-00751-9

[10] Cheong, J., Kuzucu, S., Kalkan, S., & Gunes, H. (2023). Towards gender fairness for mental health prediction. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI-23)* (pp. 5932–5940). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2023/658

[11] Sajjadian, M., Lam, R. W., Milev, R., Rotzinger, S., Frey, B. N., Soares, C. N., Parikh, S. V., Foster, J. A., Turecki, G., Müller, D. J., Taylor, V. H., Kennedy, S. H., & Uher, R. (2021). Machine learning in the prediction of depression treatment outcomes: A systematic review and meta-analysis. *Psychological Medicine, 51*(16), 2742–2751. https://doi.org/10.1017/S0033291721003871

[12] LoParo, D., Dunlop, B. W., Nemeroff, C. B., Mayberg, H. S., & Craighead, W. E. (2025). Prediction of individual patient outcomes to psychotherapy vs medication for major depression. *npj Mental Health Research, 4*, Article 10. https://doi.org/10.1038/s44184-025-00119-9

[13] Xu, X., Zhang, H., Sefidgar, Y., Ren, Y., Liu, X., Seo, W., Brown, J., Kuehn, K., Merrill, M., Nurius, P., Patel, S., Althoff, T., Morris, M. E., Riskin, E., Mankoff, J., & Dey, A. K. (2022). GLOBEM dataset: Multi-year datasets for longitudinal human behavior modeling generalization. In *Advances in Neural Information Processing Systems* (Vol. 35, Datasets and Benchmarks Track). https://doi.org/10.48550/arXiv.2211.02733

[14] Dang, V. N., Cascarano, A., Mulder, R. H., Cecil, C., Zuluaga, M. A., Hernández-González, J., & Lekadir, K. (2024). Fairness and bias correction in machine learning for depression prediction across four study populations. *Scientific Reports, 14*(1). https://doi.org/10.1038/s41598-024-58427-7

[15] Balliu, B., Douglas, C., Seok, D., Shenhav, L., Wu, Y., Chatzopoulou, D., Kaiser, W., Chen, V., Kim, J., Deverasetty, S., Arnaudova, I., Gibbons, R., Congdon, E., Craske, M. G., Freimer, N., Halperin, E., Sankararaman, S., & Flint, J. (2024). Personalized mood prediction from patterns of behavior collected with smartphones. *npj Digital Medicine,* 7, Article 49. https://doi.org/10.1038/s41746-024-01035-6

[16] Terhorst, Y., Messner, E.-M., Asare, K. O., Montag, C., Kannen, C., & Baumeister, H. (2025). Investigating smartphone-based sensing features for depression severity prediction: Observation study. *Journal of Medical Internet Research, 27*, Article e55308. https://doi.org/10.2196/55308

[17] Vairavan, S., Rashidisabet, H., Li, Q. S., Ness, S., Morrison, R. L., Soares, C. N., Uher, R., Frey, B. N., Lam, R. W., Kennedy, S. H., Trivedi, M., Drevets, W. C., & Narayan, V. A. (2023). Personalized relapse prediction in patients with major depressive disorder using digital biomarkers. *Scientific Reports, 13*, Article 18596. https://doi.org/10.1038/s41598-023-44592-8

[18] Abd-alrazaq, A., AlSaad, R., Harfouche, M., Aziz, S., Ahmed, A., Damseh, R., & Sheikh, J. (2023). Wearable artificial intelligence for detecting anxiety: Systematic review and meta-analysis. *Journal of Medical Internet Research, 25*, Article e48754. https://doi.org/10.2196/48754

[19] Ahmed, I., Brahmacharimayum, A., Ali, R. H., Khan, T. A., & Ahmad, M. O. (2025). Explainable AI for depression detection and severity classification from activity data: Development and evaluation study of an interpretable framework. *JMIR Mental Health, 12*, Article e72038. https://doi.org/10.2196/72038

[20] Kim, A. Y., Jang, E. H., Lee, S.-H., Choi, K.-Y., Park, J. G., & Shin, H.-C. (2023). Automatic depression detection using smartphone-based text-dependent speech signals: Deep convolutional neural network approach. *Journal of Medical Internet Research, 25*, Article e34474. https://doi.org/10.2196/34474

[21] Riad, R., Denais, M., de Gennes, M., Lesage, A., Oustric, V., Cao, X. N., Mouchabac, S., & Bourla, A. (2024). Automated speech analysis for risk detection of depression, anxiety, insomnia, and fatigue: Algorithm development and validation study. *Journal of Medical Internet Research, 26*, Article e58572. https://doi.org/10.2196/58572

[22] Song, M., Yang, Z., Triantafyllopoulos, A., Zhang, Z., Nan, Z., Tang, M., Takeuchi, H., Nakamura, T., Kishi, A., Ishizawa, T., Yoshiuchi, K., Schuller, B., & Yamamoto, Y. (2024). Empowering mental health monitoring using a macro-micro personalization framework for multimodal-multitask learning: Descriptive study. *JMIR Mental Health, 11*, Article e59512. https://doi.org/10.2196/59512

[23] Weber, J., Weber, M., & Lopez Alcaraz, J. M. (2025). Depression diagnosis from patient interviews using multimodal machine learning. *Frontiers in Psychiatry, 16*, Article 1694762. https://doi.org/10.3389/fpsyt.2025.1694762

[24] Jin, N., Ye, R., & Li, P. (2025). Diagnosis of depression based on facial multimodal data. *Frontiers in Psychiatry, 16*, Article 1508772. https://doi.org/10.3389/fpsyt.2025.1508772

[25] Li, Y., Yang, X., Zhao, M., Wang, J., Yao, Y., Qian, W., & Qi, S. (2025). Predicting depression by using a novel deep learning model and video-audio-text multimodal data. *Frontiers in Psychiatry, 16*, Article 1602650. https://doi.org/10.3389/fpsyt.2025.1602650

[26] Sharma, S. K., Alutaibi, A. I., Khan, A. R., Tejani, G. G., Ahmad, F., & Mousavirad, S. J. (2025). Early detection of mental health disorders using machine learning models using behavioral and voice data analysis. *Scientific Reports, 15*, Article 16518. https://doi.org/10.1038/s41598-025-00386-8

[27] Shin, D., Kim, H., Lee, S., Cho, Y., & Jung, W. (2024). Using large language models to detect depression from user-generated diary text data as a novel approach in digital mental health screening: Instrument validation study. *Journal of Medical Internet Research, 26*, Article e54617. https://doi.org/10.2196/54617

[28] Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., & Li, K. (2024). Large language models for mental health applications: Systematic review. *JMIR Mental Health, 11*, Article e57400. https://doi.org/10.2196/57400

[29] Ntam, V. A., Huebner, T., Steffens, M., & Scholl, C. (2025). Machine learning approaches in the therapeutic outcome prediction in major depressive disorder: A systematic review. *Frontiers in Psychiatry, 16*, Article 1588963. https://doi.org/10.3389/fpsyt.2025.1588963

[30] Ortiz, A., Hintze, A., Burnett, R., Gonzalez-Torres, C., Unger, S., Yang, D., Miao, J., Alda, M., & Mulsant, B. H. (2022). Identifying patient-specific behaviors to understand illness trajectories and predict relapses in bipolar disorder using passive sensing and deep anomaly detection: Protocol for a contactless cohort study. *BMC Psychiatry, 22*, Article 288. https://doi.org/10.1186/s12888-022-03923-1

[31] Pan, Y., Wang, P., Xue, B., Liu, Y., Shen, X., Wang, S., & Wang, X. (2025). Machine learning for the diagnosis accuracy of bipolar disorder: A systematic review and meta-analysis. *Frontiers in Psychiatry, 15*, Article 1515549. https://doi.org/10.3389/fpsyt.2024.1515549

[32] Khalil, A., Hussain, M., Majeed, M. K., Hamza, A., Ali, A., Ajaz, K., Siddiqui, M. H. S., & Abbasi, M. D. (2025). Artificial intelligence in neuro-oncology: Integrating advanced machine learning techniques for accurate and early detection of brain tumors through MRI imaging. *Spectrum of Engineering Sciences, 3*(8), 413–435. https://thesesjournal.com/index.php/1/article/view/826

[33] Khalil, A., Hussain, H. A. W., Khan, A. H., Majeed, M. K., Abbasi, M. D., Siddiqui, M. H. S., & Baig, A. K. K. (2025). Transforming heart transplantation with AI: Deep neural networks for predictive analytics and real-time monitoring in clinical decision support systems. *Multidisciplinary Surgical Research Annals, 3*.

[34] Shakeel, K., Hussain, S. S., Khalid, M., Ghaffar, F., Ali, I., Saif, Z., et al. (2026). A unified benchmark of statistical, machine learning, and deep learning approaches for S&P 500 index forecasting. *Spectrum of Engineering Sciences, 4*(3), 597–619.

[35] Mir, S. Z., Majeed, M. K., Siddiqui, M. H. S., Khawer, S. K., Younus, I., Shah, S. R., et al. (2025). Integrating AI and IoT in solar power systems for improved monitoring and predictive maintenance. *Spectrum of Engineering Sciences*, 971–982.

[36] Haq, R. U., Aman, F., Majeed, M. A., Raza, S., Khan, A., Hussain, R., & Majeed, M. K. (2025). Developing edge computing solutions for IoT devices to reduce latency and enhance real-time decision-making. *Spectrum of Engineering Sciences, 3*(4), 961–970. https://thesesjournal.com/index.php/1/article/view/323